Information Technology and Libraries, v10 n2 p87-98 Jun 1991
ISSN: 0730-9295
http://www.lita.org

Techniques to Improve Subject Retrieval in Online Catalogs: Flexible Access to Elements in the Bibliographic Record

Tschera Harkness Connell is a doctoral student, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

A paragraph description of "what the book is about" taken from Book Review Digest is used to evaluate information on the bibliographic record. It is first determined to what extent book descriptions match either subject headings or keywords in the title. Segments of the bibliographic record are then examined to determine their potential for retrieving the book described. The combination of approaches used to simulate manipulation of the data in the record increased recall in the sample by 20%. Keyword matching in the personal name and corporate name subject fields is a way to increase both precision and recall.

Providing adequate subject access is one of the most important challenges facing librarians today. Online catalog use studies have shown that users search by subject at least as frequently as they do by title or author.[1] However, the difficulties of providing good subject access are numerous. There is a great deal of subjectivity involved in determining what a book is about. The problem is compounded when predicting how a patron will ask for the book. Variances in language, terminology, semantics, and point of view may cause the same book to be described differently by indexer and user.

There are at least three approaches to solving the problem of subject access in an online environment: (1) to enhance the content of the bibliographic record by augmenting it with additional information, (2) to educate the user about the strengths and the limitations of the system, and (3) to develop the interface between the content and users so that the chances of a user's input matching the content of the system will be increased. These approaches are not mutually exclusive but are a convenient way of isolating various facets of system design for further study. This study is concerned primarily with the third approach.

Interface design can include systems features for manipulating user input such as automatic right truncation, spelling checkers, and simultaneous singular/plural retrieval. Interface design also can include features that manipulate the grammar and punctuation of designated fields in the record, or that search parts of records that have not traditionally been used for direct access.

Studies have shown that the match rate between user terms and catalog subject headings ranges widely from a low of 14% to a high of 58%[2]. However, user terms are frequently brief and may not represent topics within the scope of the database. In this study, a paragraph description of "what the book is about" is taken from *Book Review Digest* and used to evaluate the information on the bibliographic record. The degree to which book descriptions match assigned subject headings indicates the chances that a user will be able to find the book if the user makes a request for "what the book is about."

The study proceeds in two phases. The first phase determines to what extent book descriptions match either subject headings or keywords in the title. The second phase examines segments of the bibliographic record that presently are not widely used in subject retrieval in order

to determine their potential for retrieving the book described. This second phase is performed only on those books that did not match in the first phase. The analyses concentrate on permutations and/or segmentations of the Library of Congress (LC) subject headings as found on Library of Congress records in the Online Computer Library Center (OCLC) database.

STATEMENT OF THE PROBLEM

Manual catalogs constrain users by the linear, alphabetic arrangement of subject headings filed in the catalog. It is possible in manual catalogs to provide multiple entries and thus have multiword access, but the size of most library collections precludes an exhaustive listing of potential access points. Computers make it easier both to increase the quantity of information in the record and to access more points in the record. However, it is important to understand the ramifications of this computer potential before massive modifications to online catalogs are made. Past experience with major changes in systems of bibliographic organization in libraries (new cataloging codes, new shelf classification schemes) have raised questions as to whether the "improvements" have warranted the costs involved. The library faces the costs of time and resources to make the changes. The user has to grapple with using multiple systems existing side by side.

Therefore, before we increase access points to the catalogs of our collections it is important to know which enhancements are most likely to increase retrieval of relevant information. More access is not necessarily better access. Some modifications may not warrant the expense of time and personnel to effect the change; others may increase recall, meaning the number of items retrieved, but at the same time decrease the precision ratio, that is, the proportion of items that are relevant to the request.

Grammar and syntax in the *Library of Congress Subject Headings (LCSH)* are often seen as a further hindrance to user access. It may be possible to work around this access problem by using the flexibility of the computer for multiple access without having to make the corresponding changes in card or book catalogs. This study looks at the results of some of these more flexible means of access.

LITERATURE REVIEW

Enhancing the Content through Augmentation

Enhancing the content of the record can be accomplished by adding information to the record and by using the information already available to better advantage. Several researchers have suggested adding more information to the catalog record. The most extensive study of this approach was the Subject Access Project (SAP) performed under the direction of Atherton.[3] A monographic database was created by augmenting the subject headings of 2400 MARC (MAchine Readable Cataloging) records with words and phrases from the book's index and/or table of contents. Searches were then performed in the augmented database, BOOKS, and compared with the same searches performed on records that had not been augmented. Online searches in BOOKS took less time and retrieved more relevant items than the same searches in the MARC database. Although the precision ratio for BOOKS was poor, it was no worse than for searches in the unaugmented MARC database: both systems produced two to three nonrelevant items for every one relevant item retrieved. Despite the problems of precision, SAP demonstrated that improvement in subject recall could be obtained by adding selected content and index information to the bibliographic record. The costs of adding and storing the additional information were relatively low when compared to traditional subject analysis costs.

Adding a greater variety of authorized terms to the subject headings list is an indirect way of augmenting the database. LC's policy of literary warrant means that a subject heading term is created only if the Library processes a book that requires it. Therefore, libraries with specialized collections may find *LCSH* inadequate for their needs. The Library of Congress recognizes this problem. One of the purposes of the publication of the Library of Congress' *Subject Cataloging Manual* is to aid "practicing subject catalogers wishing to assign subject headings in the spirit of LC's own policies and practices."[4] The Library of Congress has encouraged individuals wishing to submit headings for *LCSH* to write for instructions and blank subject authority worksheets.[5]

Another way of adding information to the record is to use classification to enhance subject retrieval. Using the OCLC MARC records as a research base, Markey and Demeyer have tested the retrieval potential of the Dewey Decimal Classification (DDC).[6] Subject terms in the *DDC Schedules and Relative Index,* hierarchical arrays of related terms in the schedules, and class numbers in the schedules and index were all used as searchers' tools for subject access and browsing. DDC numbers on the bibliographic record were indexed to provide potential access points. Markey and Demeyer demonstrated that, used in combination with *LCSH* and title keywords, DDC could increase recall and bring up relevant items that would not be retrieved by any of the other fields on the record.

Chan has examined the theoretical issues related to whether the Library of Congress Classification (LCC) can be used to enhance subject retrieval in an online catalog while Williamson is working on a project to test the potential of using LCC for subject retrieval.[7,8] Huestis has reviewed strategies for overcoming the major problems involved in the use of LCC as an access point in online catalogs.[9]

Enhancing Access through Improvement *of LCSH*

Improving access to the bibliographic record can also be accomplished by using the information already available to better advantage. Although the MARC record provides the means for using several sources for subject headings, including headings that are locally assigned, the primary source for subject headings in the United States is the *Library of Congress Subject Headings* list *(LCSH).* Improving *LCSH* is a way to improve the bibliographic record.

The *Library of Congress Subject Headings* list has been extensively criticized during the seventy-plus years of its existence.[10] Outside the scope of this study, but of major concern, have been issues of terminology. One criticism has been the inability of the system to keep up with current terminology, especially in rapidly evolving fields of knowledge. Another has been the Male- /Anglo- /Western-/ Judeo-Christian bias of the terminology used. The structure and grammar of *LCSH* have also been major areas of concern. The *LCSH* reference structure consists of *see* references from terms not used, to the headings that are used and *see also* references from a broader heading to a narrower term, or between any two headings that are related other than hierarchically. The 1983 Library of Congress entry vocabulary project was one experimental effort to increase the coverage of *see* references in the *Library of Congress Subject Headings* list**.**[11]

The immense size and rapid growth of the Library of Congress collections have meant that inconsistencies and anomalies have developed in the list. In 1970, Harris showed that some of the headings which cause problems in terminology are leftovers from earlier lists, and *LCSH* cross-references were shown to be inconsistent and incomplete.[12]

In 1972, a study conducted by Sinkankas pointed out the need for an improvement in the entire reference structure of *LCSH.* Sinkankas tested the *see also* references of *LCSH* to determine if the hierarchy of references would guide the user through a subject until all aspects of the subject

had been exposed. He concluded that "[t]he syndetic structure does not perform any guiding function at all. It connects terms, but the connection may not be considered a classification."[13] In light of this study, the recent decision of the Library of Congress to adopt the very precisely defined terminology of thesaurus construction to show relationships in the LC subject headings list is misleading. Dykstra, in particular, criticizes the new terminology of *LCSH.*[14] "Broader term," "narrower term," and "related term" are precisely defined to show very specific hierarchical relationships in tightly structured vocabularies. The Sinkankas study shows that the syndetic reference structure of *LCSH* is not tightly controlled.

There have been many studies that have dealt with the structure and grammar of Library of Congress subject Headings. Steinweg studied headings in order to determine how common marks of punctuation (comma, parentheses, hyphen, apostrophe, and period) are used and if their usage was consistent and predictable.[15]

Wepsiec considered the grammar of the headings from a different point of view. By grouping existing headings into twenty-two syntactic types, he found that some semantic types were expressed by more than one syntactic type. For example, the headings "Religion and sociology" and "Hospitals—Sociological aspects" both present the sociological perspective of the focal noun (religion in the first case, hospitals in the second). Wepsiec suggests that the user would be better served if the two headings were formulated using the same syntax—the noun qualified by a subdivision. He concluded that seven syntactic types could be eliminated without loss of specificity of the headings.[16]

Dailey also performed an extensive evaluation of the grammar of Library of Congress subject headings.[17] He developed fifteen syntactical rules for heading formation in which he advocated using punctuation consistently and uniquely in formulating headings.

Understanding the structure and grammar of *LCSH* is important in order to make effective improvements in the list itself and to use computer systems to access the existing headings effectively in different ways. Both can be considered enhancements to the content of the bibliographic record.

Enhancing Access through Computer Design

Mischo, in two studies performed at Iowa State University, took the approach of using a computer to simulate changes in the headings. He explored the possibility of increasing access to the bibliographic record through the use of computer-assisted indexing.[18] He later experimented with the same techniques using the Online Union Catalog of OCLC. Because of the tremendous requirements for machine space to invert files needed for Boolean searching, an algorithm was developed for the rotation of significant words in a subject heading as a way of increasing subject access points. Mischo simulated "Boolean search capability over subject heading words by building precoordinated term combinations into the derived or phrase index keys."[19] The ability to search title keywords was judged to be an important integral component of the project. The fact that these approaches provided improved subject retrieval over unmodified *LCSH* is not surprising in view of the fact that a greater number of access points (an estimated fifteen per title) is provided. Increasing the potential access by adding access points was the purpose of the experiments.

Lester, in her doctoral research on improving subject access in online catalogs, used the approach of developing the interface to increase the chances of a user's input matching the content of the system. User terms collected from transaction logs at Northwestern University were compared with *LCSH* headings to determine the degree of match success. Then twenty-two systems features for manipulating user input were applied to the "match failures" (those terms that

did not exactly match *LCSH* headings) to determine how well modifications improved the rate of match success.[20]

In many ways the present study is patterned after the methodology used by Lester. However, there are two important differences in approach between the two studies. Lester's research approaches the problem of improving subject access primarily by modifying user input. The present study approaches the problem by simulating modifications of the content in the bibliographic record. There are infinite ways a user can ask for information. The studies into the structure and grammar of *LCSH* indicate that there is a finite number of syntactic patterns for Library of Congress subject headings. This project examines the potential benefits of manipulating four of those patterns with and without using keywords in the title. Therefore, the focus is not on user success but on potential user success. The use of an independent description of "what the book is about" addresses the issue of whether the indexing reflects the content of the book.

METHODOLOGY

The *Book Review Digest (BRD)* was selected as the source of independent summaries of each book's content. The *BRD* "provides excerpts of and citations to reviews of current fiction and nonfiction in the English language."[21] It also includes a summary paragraph for each entry on "what the book is about."

Because juvenile and young adult materials are not a collection priority of LC, the terminology of *LCSH* is often inadequate for these materials. Since this study examines the effect of modifying traditional access to library of Congress subject headings, juvenile and young adult titles are eliminated.

The books listed in the 1987 *Book Review Digest* (excluding juvenile and young adult titles) provide the population for the study. A preliminary examination of 20 randomly chosen pages of the *BRD* produced descriptions of 66 books. The purpose of this preliminary sample was to gather information that would help determine the size of the sample required for the final study. Twenty-one percent of the entries were juvenile or young adult titles and therefore were eliminated. The remaining books were then examined to determine the proportion of entries whose *BRD* descriptions match main headings of subject headings, or keywords in the title proper. Seventy percent produced a match.

Considering this initial match of 70% in a preliminary examination of *BRD,* the expected proportion of the population that would fail to match on main headings of the subject and on keywords in the title, but succeed on any one of the final tests, is estimated to be 20% or less. The 20% figure is based on an assumption that some but not all of the 30% that failed to match on the main heading of the subject and on keywords in the title will match on one of the final tests, and that 20% is at the upper end of the percentage range that might be expected to match. To achieve a statistical precision of 2.5% in the estimated proportion (at the 95% confidence level), the required sample size is: $n = (1.96/.025)^2 (.2) (.8) = 983$.

Twelve hundred entries were randomly selected from the 1987 *BRD* to ensure that after juvenile and young adult titles books were eliminated, at least 983 other titles would remain (983 is approximately .79 X 1,200).

To obtain the random sample, 430 pages were photocopied. The 1987 *BRD* is 2,064 pages long which means that the sample size is approximately 21% (430/2064) of the entire population. The sample was chosen by dividing the *BRD* into 6 equal sections and then photocopying the first 71 pages of 2 sections, and the first 72 pages of each of the remaining 4 sections. The 430 pages produced 1,297 full *BRD* entries. Of those 1,297 entries, 266 (21%) were eliminated because they

are juvenile or young adult titles. Eight additional entries were eliminated because the descriptions for what the books were about appeared in an earlier volume of the *BRD*. The 1,297 entries yielded a sample of 1,023 book descriptions that are used in this study.

After the 1,023 books were identified, each title was searched for Library of Congress cataloging in OCLC. As each of the 1023 titles was found, a paper print was made of the OCLC control number, author, title, and all Library of Congress subject headings.

Phase 1: Match by Subject or Keywords in the Title Proper

In manual catalogs, a user will find the main subject heading and then distinguish between aspects of the topic by using the subdivisions assigned. Computer catalogs can be designed to search either the entire field or just the first element (main heading). Most online catalogs also have the capacity for searching keywords in the title. The first phase of this study determines a match rate between the *BRD* book description and the combined elements of the main heading (sub-field ± a) of Library of Congress subject headings or *LCSH* recommended cross-references, and keywords in the title proper (245 field, subfield ± a). ("Title proper" is defined as the main title including alternative title, but excluding parallel titles and other title information.) A subject heading match occurs when a term or phrase in the book description is exactly the same when compared from left to right, letter for letter, (excluding capitalization, punctuation, and birth/death dates of persons) as the term or phrase in the subject heading. A keyword match occurs when a term in the book description is exactly the same when compared from left to right, letter for letter, (excluding capitalization, punctuation, and birth/death dates of persons) as the term in the title. For keyword matches, the order of terms in the book description and in the bibliographic record do not need to be the same. Fourteen words—stop words—are excluded from comparison: a, an, and, at, by, for, from, how, in, of, on, the, to, with.

This first phase of the project is accomplished in three steps. The book descriptions, excluding stop words, are initially compared against the first element of each subject field (subfield ± a). In MARC records, subject headings appear in the 600 (6XX) numbered fields. All types of subject headings—personal name (600), corporate name (610), conference or meeting name (611), uniform title (630), topical (650), and geographic (651) — are compared.

The second step compares the book descriptions with the 10th edition of the *Library of Congress Subject Headings* (1986). A match on a reference from terminology not used to a heading that had been assigned to the book under consideration is counted as a subject match. It is not uncommon to find libraries making the recommended *see* references prescribed in *LCSH*, even when no other kinds of references are made. It is for this reason that evaluating matches on *see* references is included. It was also the desire of this researcher to determine the degree to which the use of the recommended *see* references increases the match rate, before comparisons on keywords in the title are made. However, it is recognized that many libraries do not make any references in their catalogs. For this reason the results of subject matches achieved through *see* references are kept separate from the results of direct subject matches. All books which have a subject match were removed from further consideration.

The third step compares the remaining book descriptions (those that had not matched on subject) with the keywords in the title proper (± a) (see table 1).

Table 1. Matches between Description of Book Content and LC Subject Headings and Keywords in the Title (sample size = 1,023 books)

Subject (≠ a)	Subject (x–ref)	Title (keyword)	Total Matches	No–Match
365	37	284	686	337
(35.7%)	(3.6%)	(27.8%)	(67%)	(33%)

Phase 2: Match on a Simulated Manipulation of the Bibliographic Record

The second phase of the project involves a series of five tests on the 337 unmatched items from the first phase. All five tests were performed on each of the 337 titles. The tests measure the improvement in access to the existing MARC record obtained by modifying the way information in the record is accessed. All of the five methods of access can be achieved through interface design; three of the tests simulate modifying the grammar and syntax of the subject headings themselves. The purpose of this phase of the study is to determine to what extent these additional access points will increase the chances that books that did not match by subject or keywords in the tide will now produce a match.

Four tests compare the descriptions of what the book is about with the Library of Congress subject headings as they appear in current MARC records. The fifth test compares the description against keywords in any portion of the title other than the title proper. The same stop words used in phase 1 are used also for this phase of the study. The results of these tests show the amount of improvement in match results for each modification made. The tests performed are comparisons between descriptions of what each book is about and:

1. subdivisions of the main subject fields.
Example: The phrase "data processing" will
produce a match with the heading, EDUCATION—DATA PROCESSING.
2. inverted subject headings changed to direct order. This test was divided into a comparison of name headings (inversions, name) and topical headings (inversions, nonname).
Examples: The phrase "sociology of knowledge" will produce a match with the subject heading, KNOWLEDGE, SOCIOLOGY OF. The name "Anthony Eden" will match with the subject heading: EDEN, ANTHONY.
3. the principal term or phrase in a subject heading involving a parenthetical qualifier. Words or phrases in the book description were compared with the principal element of the subject heading; that is, up to the first parenthesis. In a manual catalog, the parenthetical qualifier helps the user to distinguish between two headings that would appear the same without the qualifier. This test simulates the manual environment in that it makes the initial comparison on the unqualified heading.
Examples: "Alabama" will match with the subject heading, ALABAMA (MUSICALGROUP). "Power" will match with the subject heading, POWER (SOCIAL SCIENCES).
4. keywords in the subject headings. This test was divided into a comparison of keywords involving proper names (keywords, name), and keywords not involving names (keywords, nonname).
Examples: "Bowie" and "Kuhn" will both match the subject heading KUHN, BOWIE,

1926- "1926" will not match because birth and death rates are excluded from the matching process. "Power," "social," and "sciences" will all match the subject heading POWER (SOCIAL SCIENCES).

5. keywords in parallel tides, and/or other title information (field 245 ± b).

Example: The words "funny," "pro," "business," and "football" will each match with the other title information of *First down and a billion: the funny business of pro football.*

A match is defined the same as for phase 1: the term or phrase in the book description must be exactly the same when compared from left to right, letter for letter, (excluding capitalization, punctuation, and birth/death dates for persons) as the term on the record. For tests involving keyword matching, individual terms do not have to be in the same order in the book description as in the bibliographic record.

Data for each of the five tests are recorded from three points of view. First, the total number of match occurrences for each test is recorded. For example, if a term in the book description produces a match in three of the five tests, then three matches are recorded: one match for each test. Considering only questions of recall, these data help answer the question: If we can only add one of the tested methods of access, which one? The 337 book descriptions, when compared with all five modifications to the record (the five tests), produce a total of 405 hits. Table 2 gives a summary of these data.

Table 2. Total Number of Matches for Each Test (Presented in rank order)(n = 405 [total no. matches for 337 book descriptions]) (π = population proportion)

	No. Matches	Confidence Interval*
Keywords	169	.3704 ≤ π ≤ .4659
Keyword (nonname)	127	.2703 ≤ π ≤ .3603
Keyword (name)	42	.0776 ≤ π ≤ .1372
Title (245 | b)	136	.2915 ≤ π ≤ .3832
6XX subfields	67	.1324 ≤ π ≤ .2047
Inversions	17	.0264 ≤ π ≤ .0662
Inversions (name)	14	.0207 ≤ π ≤ .0572
Inversions (nonname)	3	.0025 ≤ π ≤ .0215
Parenthetical Qualifier	16	.0245 ≤ π ≤ .0632
No-Match	130	

*Formula for determining confidence interval taken from Glass and Hipkins, *Statistical Methods in Education and Psychology*, 2d ed. (Englewood Cliffs, N.J.: Prentice-Hall) p.280.

The second approach records unique matches. That is, a match is recorded for a test if it is the only test which produces a match. For example, if one book description matches in three of the five tests, no match is recorded. However, if a match occurs on the fourth test only, then the match is recorded for the fourth test. Considering only recall, this approach answers the question: Given all five methods of access, which one can we eliminate with the least harm? Table 3 represents this approach.

The third approach to the data is a variation on the second. The difference between the two is that the fifth test, determining the number of matches on keywords in the sub-field ± b of the

tide, was not used in the analysis. Therefore, a book description that was not recorded for the second approach because it matched on tests 3 and 5, would be recorded for the third approach, because it matched only on test 3. This point of view answers the question: Which of the four modifications to subject headings has the least (or most) potential to produce matches with books that are not matched by any of the other methods tried? Table 4 is a summary of this approach.

Table 3. Tests that Produced Unique Matches: Title Subfield ≠ b Included (Presented in rank order) (Sample size = 1,023) (π = population proportion)

	No. Matches	Confidence Interval		
Title (245 ≠ b)	40	.0288	≤ π ≤	.0528
Total Keywords	34	.0239	≤ π ≤	.0461
Keyword (name)	8	.0040	≤ π ≤	.0154
Keyword (nonname)	26	.0174	≤ π ≤	.0370
6XX subfield	19	.0119	≤ π ≤	.0288
Parenthetical Qualifier	3	.0010	≤ π ≤	.0086
Total Inversions	1	.0002	≤ π ≤	.0055
Inversions (name)	1	.0002	≤ π ≤	.0055
Inversions (nonname)	0	.0000	≤ π ≤	.0037
No-Match	130	.1080	≤ π ≤	.1489
No-Match (Fiction)	57	.0433	≤ π ≤	.0715

Table 4. Tests that Produced Unique Matches: Title Subfield ≠ b Excluded (Presented in rank order) (Sample size = 1,023) (π = population proportion)

	No. Matches	Confidence Interval		
Total Keywords	91	.0730	≤ π ≤	.1080
Keyword (name)	24	.0158	≤ π ≤	.0347
Keyword (nonname)	67	.0519	≤ π ≤	.0823
6XX subfield	39	.0280	≤ π ≤	.0517
Total Inversions	8	.0040	≤ π ≤	.0154
Inversions (name)	6	.0027	≤ π ≤	.0127
Inversions (nonname)	2	.0005	≤ π ≤	.0071
Parenthetical Qualifier	4	.0015	≤ π ≤	.0100
No-Match	170	.1446	≤ π ≤	.1902
No-Match (Fiction)	70	.0545	≤ π ≤	.0856

Data Analysis and Discussion

Phase 1

In the first phase of the project, exact matches on subfield ± a of the subject heading account for approximately 36% (365/1023) of the total sample. Combined with matches resulting from the use of cross references the figure is 39% (402/1023). These figures are within range of the results reported in studies that have matched user terms and Library of Congress subject headings.

Two benchmark figures for studies that evaluated library of Congress terms are the match rate of 58% in the University of Michigan card catalog study and the match rate of 40% in Lester's analysis of user terms taken from online transaction logs.[22] It is not surprising that the figure for

this study is less than the University of Michigan study because in a manual catalog there is opportunity for browsing. The human eye is much more flexible than the binary design of a computer. Therefore, in a manual catalog, user success does not depend upon an exact match. It might be reasonable to expect that this study, which made a comparison of Library of Congress subject headings with summary paragraphs, should produce a higher match rate than Lester's comparison with user terms. However, studies that look at user terms compared with subject terms used in a database are not looking at the indexing of particular books. In Lester's study a user term could match with any bibliographic record in the LC MARC database that had a matching subject assigned. The present study, by looking at the potential for finding each particular title in the sample, examines only the indexing assigned to that particular title. The population of adult titles found in *BRD* includes many fiction titles. Since it is Library of Congress policy not to provide subject headings for single works of fiction or collections of fiction by one author, there will be no match for these items.

Another factor that might contribute to the fact that the results from this phase of the study are not higher is that there were a number of instances where the summary paragraph was very vague. The following is an example:

> This volume presents abstracts from 600 journal articles. These are presented in a regional arrangement, followed by author and subject indexes and a chronology of events related to the topic. [Book title: *Global Terrorism]*

It would be very difficult to find a meaningful subject heading that would match this description. The match rate of less than 4% for the comparison of book descriptions with the *LCSH* recommended *see* references was surprisingly low. A common assumption in the library field is that providing the LC prescribed *see* references will greatly increase the chances that users will find subject material on their topic. This study does not support that assumption. The result does not mean necessarily that *see* references are unimportant, but only that the quantity of *see* references suggested by LC is not large enough to make much difference in recall.

Phase 2

This phase of the study uses the 337 book descriptions that did not match with main subject headings or the title proper during the first phase of the study.

The data (see table 2) suggest that the greatest number of matches will occur between descriptions of what the book is about and keywords in the subject fields. At a 95% confidence level, keywords in the subject fields account for between 37% and 47% of the access points. Data are fairly conclusive that keywords-subject account for more matches than keywords-title subfield; the confidence intervals are almost disjointed. (Because of the slight overlap in confidence intervals, there is a chance that keywords-title could account for more matches that keywords-subject.) Keywords in subfield ± b of the title will produce between 29% and 38% of all the access points. With respect to all other tests, however, the data indicate that matching keywords in the subject fields potentially will result in the greatest recall.

The usefulness of 6XX subfields for retrieval has been questioned on the basis that many 6XX subfields are form subdivisions (e.g., CONGRESSES, HISTORY) and therefore too general for retrieval in large databases. Because of this, the matches for the 6XX subfields were analyzed in order to determine what proportion are form subdivisions. The matches on title subfields were similarly analyzed, because in the process of data collection it became apparent that many of the title subfields indicated form. It was determined that 47.76% (32/67) of the matches in the 6XX

subfields are with terms that represent form. In the tide subfield the percentage is lower: 26.70% (36/136).

When considering ways to improve precision, searching for keywords in the personal and corporate name fields only, rather than all the subject fields, is an option to be considered. Approximately 25% (42/169) of the matches on keywords in the subject fields are on names (see table 2). Keyword name subject searching is the fourth highest group in terms of recall of the nine recorded. However, the groups that rank second (title subfield ± b) and third (6XX subfields) have a large component of matches due to form headings. Nearly 50% of the matches with the 6XX subfields are matches with form subdivisions. Although the potential increase in access through the use of keywords in the personal and corporate name subject headings is only between 8% and 14%, the benefits in increased (or at least not decreased) precision make this option worth considering.

Table 3 records those book descriptions that matched on one test and one test only. The data show that comparisons of the description of "what the book is about" with subfield ± b of the tide will match between *3%* and 5% of the book descriptions in the sample once all other techniques have failed. The potential for total recall is increased by 3-5%. Inverted subject headings and headings with parenthetical qualifiers are the least likely to increase total recall. For books that do not match by any other means, the potential success rate using these two methods in less than 1% each. Using keyword name comparisons, the potential for unique hits is less than 1.5%.

It can be expected that even by using all methods, phases 1 and 2, 11-15% of the books compared will not match. Looking at this another way, we can be 95% confident that by using all the methods of access in the study, between 85% and 89% of the book descriptions in *BRD* will produce a match. If the capability for matching on keywords in subfield ± b of the tide is removed, then the total number of potential matches is decreased by 3-5%.

Table 4 presents the results of the book descriptions that match on only one test, but only the first four tests are included. Results of comparisons of keywords in the subtitle are excluded. These data give an indication of the relative merit of each of the modifications to Library of Congress subject headings. Of the last four methods, keywords (name and non-name) comparisons are clearly the most effective method for matching books that matched by no other means. At a 95% confidence level, comparisons of summary paragraphs with keywords have the potential to match an additional 7-11% of the tides. This 7-11% is over and above any other matches made using subject headings. The next highest group is the 6XX subfield matches which range from 3 to 5%. Again, considering only recall, inversions and headings with parenthetical qualifiers are the potentially least useful modifications to be made.

This study shows that the potential match rate for descriptions of "what the book is about" with subjects and tides is between 85% and 89%. Ideally one would want the potential to be 100%. With that in mind, an analysis was performed on the books that did not match in order to determine if there were other modifications that might be made to the existing record that would increase the potential for recall. Table 5 is a summary of that analysis.

Given the Library of Congress' policy of not providing access to most works of fiction, it is not surprising that fiction accounted for the largest percentage (44%) of the nonmatches. The 11% of the nonmatches in category 3 could be retrieved by techniques of truncation. These techniques could be applied to either the bibliographic record or to user input. Categories 2, 4, and 5 are a little more difficult.

Table 5. Items That Did Not Match on any Test (Number of no-matches = 130)

Categories of Possible Reasons for No-Match	No.	%
1. Fiction titles (including collections of poems or collections of novellas	57	(43.85)
2. Works about a person or corporate body (including autobiography) where the name of the person or body was not mentioned in the book description	23	(17.69)
3. Differences in grammar (possessive form of name in description, singular/plurals, nouns/adjectives)	14	(10.77)
4. Book description more specific than subjects assigned	15	(11.54)
5. Book description more general than subjects assigned	13	(10.00)
6. Other differences in terminology	8	(6.15)
Totals	130	(100.00)

Traditional subject heading practice directs the cataloger not to assign general headings to an item that is specific. Considering category 2, this means that the heading BASEBALL PLAYERS would not be assigned to a biography of Ozzie Smith. Many of the 18% of the items did not match because the name of the person about whom the book is written was not mentioned in the *BRD* description. These items would match if general (broader) headings were assigned. In the 1970s the Library of Congress began to assign general and specific headings to certain categories of works. However, the results have not been satisfactory and the policy is being reconsidered. Users who come to the catalog looking for a general work on baseball players should be able to find what they need at the specificity they desire without having to wade through hundreds of biographies of individual players.

Categories 4 and 5 are more elusive. These cases include instances where the book summary describes just one facet of the book, or where the description is so general that it says little. Category 6 includes all other non-matches in the sample.

SUMMARY AND CONCLUSIONS

This study is concerned primarily with improving subject access in an online catalog by using the information in the bibliographic record to better advantage than can be done in a manual catalog. It has examined ways to increase subject recall by manipulating the grammar of headings and by accessing parts of fields not ordinarily accessed.

Due to the many ways that the subject content of a book can be described, it is important to provide a variety of approaches to the item. The combination of approaches used in this study increased the recall in the sample by 20% (from 67% to 87%). If recall is the only consideration, keywords in the subject fields will produce the best results. One way to increase *both* precision and recall is to provide keyword matching in the personal name and corporate name subject fields. Inverted headings and headings with parenthetical qualifiers occur infrequently. However, when they do occur, matches on these kinds of headings are likely to be precise.

Interfaces between the users and the content of a system can be designed to search in order of defined priorities. When no match occurs by using the techniques in the first phase of this study, the system can be designed to search additional fields in a specified order. Those fields that can be defined as subject-rich, such as the inverted subject headings or the headings with parenthetical qualifiers, can be searched first, before the more general fields such as the 6XX subfields are

searched.

In a sense, Carol Mandel presented the charge for this study when she cautioned that "it is important that we do *not* increase the effort and expense of record creation unless we are gaining enhancements that cannot otherwise be achieved through good online catalog design."[23] The techniques presented in this study are among those that can be used to improve subject access in online catalogs.

REFERENCES AND NOTES

1. Pauline A. Cochrane and Karen Markey, "Catalog Use Studies Since the Introduction of Online Interactive Catalogs: Impact on Design for Subject Access," *Library and Information Science Research* 5: 339 (1983); Carolyn O. Frost, "Subject Searching in an Online Catalog (Survey Conducted at the University of Houston)," *Information Technology and Libraries* 6:60-63 (Mar. 1987).
2. Marilyn A. Lester, "Coincidence of User Vocabulary and Library of Congress Subject Headings: Experiments to Improve Subject Access In Academic Library Online Catalogs" (Ph.D. diss., Univ. of Illinois, 1989), p.117-18.
3. Pauline Atherton [Cochrane], *Books Are For Use: Final Report of the Subject Access Project to the Council on Library Resources* (Syracuse, N.Y.: Syracuse University School of Information Studies, 1978).
4. Library of Congress, Office for Subject Cataloging Policy, *Subject Cataloging Manual: Subject Headings,* 3d ed., v.l (Washington, D.C.: Cataloging Distribution Service, Library of Congress, 1990), p.ii.
5. Mary K. D. Pietris, "Establishing Subject Headings in the Library of Congress" *Cataloging Service Bulletin* 41:83 (Summer 1988).
6. Karen Markey and Anh N. Demeyer, *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated Into the Subject Searching Capabilities of an Online Catalog* (Dublin, Ohio: OCLC, 1986).
7. Lois Mai Chan, "Library of Congress Classification as an Online Retrieval Tool: Potentials and Limitations," *Information Technology and Libraries* 5:181-92(Sept. 1986).
8. Nancy J. Williamson/'Classification in Online Systems—Research and Progress," in *Librarianship in Japan; [Proceedings of the] International Federation of Library Associations and Institutions 52d General Conference; 1986 August; Tokyo, Japan* (1986) (Tokyo: Japan Organizing Committee of IFLA, 1986), p.25-42.
9. Jeffrey C. Huestis, "Clustering LC Classification Numbers in an Online Catalog for Improved Browsability," *Information Technology and Libraries* 7:381-93 (1988).
10. Monika Kirtland and Pauline Cochrane, "Critical Views *of LCSH*—Library of Congress Subject Headings: A Bibliographic and Bibliometric Essay," *Cataloging & Classification Quarterly* 1:71-94 (1982).
11. "LC Subject Entry Vocabulary Project," *RTSD Newsletter* 7:66-67 (Sept.-Nov. 1982).
12. Jessica Lee Harris, *Subject Analysis: Computer Implications of Rigorous Definition* (Metu Traditional subject heading practice directs the cataloger not to assign general headings to an item that is specific. Considering category 2, this means that the heading BASEBALL PLAYERS would not be assigned to a biograchen, N.J.: Scarecrow, 1970).
13. George M. Sinkankas, *A Study in the Syndetic Structure of the Library of Congress List of Subject Headings* (Pittsburgh, Penn.: University of Pittsburgh, Graduate School of Library and Information Sciences, 1972), p.51.

14. Mary Dykstra, "LC Subject Headings Disguised as a Thesaurus," *Library Journal* 113:42-46 (Mar. 1, 1988).
15. H. Steinweg, "Punctuation in the Library of Congress Subject Headings," *Library Resources & Technical Services* 22:145-53 (Spring 1978).
16. Jan Wepsiec, "Language of the Library of Congress Subject Headings Pertaining to Society," *Library Resources & Technical Services* 25:196-203 (Apr. 1981).
17. Jay E. Dailey, "The Grammar of Subject Headings: A Formulation of Rules for Subject Headings Based on a Syntactical and Morphological Analysis of the Library of Congress List," in Pauline A. Cochrane, *Improving LCSH For Use In Online Catalogs* (Littleton, Colo.: Libraries Unlimited, 1986), p.159-64.
18. William H. Mischo, "Expanded Subject Access to Reference Collection Materials," *Journal of Library Automation* 12:338-354 (Dec. 1979).
19. William H. Mischo, "Subject Retrieval Function Based on Computer-Manipulated Library of Congress Subject Headings," in *Information Interaction: Proceedings of the 45th ASIS Annual Meeting, Columbus, Ohio, October 17-21,1982* (White Plains, N.Y.: Published for the American Society for Information Science by Knowledge Industry Publications, 1982), p. 197.
20. Lester, *Coincidence of User Vocabulary.*
21. Martha T. Mooney, ed., *Book Review Digest* (March 1987 to February 1988 inclusive) (New York: Wilson, 1988), prefatory note.
22. Renata Tagliacozzo and others, *Patterns of Search in Library Catalogs: An Empirical Study*, in Manfred Kochen, *Integrative Mechanisms in Literature Growth*, v.2, Part IV, Final Report to the National Science Foundation (Ann Arbor, Mich.: Mental Health Research Institute, University of Michigan, 1970).
23. Carol A. Mandel, "Enriching the Library Catalog Record for Subject Access," in Pauline Cochrane, *Improving LCSH for Use in Online Catalogs* (Littleton, Colo.: Libraries Unlimited, 1986), p.233.